DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 17 | Issue 1 Article 8

2018

Optimal Model Selection for Truncated Data among Non-Nested Competitive Models

Parisa Torkaman
Malayer University, Malayer, Iran, p.torkaman@malayeru.ac.ir

Recommended Citation
Torkaman, P. (2018). Optimal Model Selection for Truncated Data among Non-Nested Competitive Models. Journal of Modern Applied Statistical Methods, 17(1), eP2379. doi: 10.22237/jmasm/1525132980

Journal of Modern Applied Statistical Methods
May 2018, Vol. 17, No. 1, eP2379
doi: 10.22237/jmasm/1525132980

ISSN 1538 − 9472

Optimal Model Selection for Truncated Data among Non-Nested Competitive Models

Parisa Torkaman
Malayer University
Malayer, Iran

Selecting a model for incomplete data is an important issue. Truncated data is an example of incomplete data, which sometimes occurs due to inherent limitations. The maximum likelihood estimator features and its asymptotic distribution are studied, and a test statistic among non-nested competitive model of incomplete data is presented, which can select an appropriate model close to the true model. This close-to-true model under the null hypothesis of the equivalency of two competitive models against alternative hypothesis is selected.

Keywords: Kullback–Leibler information criteria, non-nested competitive model, truncated data

Introduction

To draw an inference from a population, selecting an appropriate model is critical. The goal is to identify an optimal model from some competitive models by observing the population and testing the related hypotheses. Model selection and hypothesis testing for the purpose of complete data has been widely studied (Cox, 1962; Vuong, 1989). Because complete data is rare, choosing an appropriate and therefore optimal model for incomplete data is important.

Considered as an example of truncated data, incomplete data exists due to inherent limitations. This truncation phenomenon in statistical distribution has become evident by close observation, and is used extensively in such sciences as astronomy, reliability, medicine, and economics. A manufacturing process, for example, cannot produce parts having a negative life; in a call center, excessively

doi: 10.22237/jmasm/1525132980 | Accepted: October 26, 2017; Published: June 7, 2018.
Correspondence: Parisa Torkaman, p.torkaman@malayeru.ac.ir

long conversations are terminated. Therefore, a portion of potential data with a wide range of values is not applicable.

In the course of selecting a true density model, some criteria are referred to when the true density model is unknown. For instance, Kullback-Leibler information (Kullback & Leibler, 1951) is a model selection criterion based on the hazard function. It determines how a model diverges from competitive models. It is first divided into two parts, and the second part predicts the true density of the data. Akaike (Akaike, 1973) introduced unbiased estimators of Akaike information for this risk, and also generalized it. Cavanaugh (1999) obtained Akaike's estimator based on the symmetric Kullback-Leibler, and found it efficient in a survey of divergence of the true model from competitive models. Vuong (1989) proposed a model selection test based on the Kullback-Leibler information, used to check the closeness of the competitive models to the true model. Cox (Cox, 1961, 1962) offered a method with a generalized likelihood ratio test for a non-nested model so that the null hypothesis included the true density of the data.

In the case of censored data, Shimodaira (1994) showed that Kullback-Leibler criterion based on both the observed and censored data is better than studying the observed data alone. In addition, in comparison with the censored data, the divergence criterion for the complete data is more sensitive. Bhattacharyya (1985) revealed that the maximum likelihood estimator for the censored data of type II is convergent to the true value parameter in probability, and that it has an asymptotic distribution. Bardley and Vahe (1999) developed nonparametric methods for testing and estimating doubly truncated data. Lominashivli and Patsatsia (2013) studied the estimation of the parameters of the exponential distribution, truncated from two sides based on the maximum likelihood method, and offered unique solutions for obtaining it.

The current study deals with the maximum likelihood estimator for truncated data, which is under-explored in the related literature. To select an appropriate and optimal model from some competitive models for truncated and incomplete data, a statistical test was conducted to determine how close the true model was to the non-nested competitive models. In the current study, a truncated distribution from both sides is placed at the interval (a, b).

Theory of Truncated Models and Results

Consider $X_1, \ldots, X_n$ as an independent random sample with the same distribution and true unknown density function $h(x)$. Also consider the competitive model

$f_\theta = \{ f_\theta(x); \theta \in \Theta \subset R^p \}$ with the unknown parameter θ, which is a member of the parameter space Θ and dimension p. A disadvantage of using $f_\theta(x)$ density instead of true density $h(x)$ is defined as $\log \frac{h(x)}{f_\theta(x)}$. The expectation of this loss function under the true model data is called the Kullback-Liebler information criteria and is shown as follows:

$$KL(h, f_\theta) = E_h\left[\log \frac{h(X)}{f_\theta(X)}\right] = E_h\left[\log h(X)\right] - E_h\left[\log f_\theta(X)\right], \qquad (1)$$

If $h(x) = f_\theta(x)$ then $KL(h, f_\theta) = 0$. Consider the two competitive models f_θ and $g_\beta = \{ g_\beta(x) : \beta \in B \subset R^q \}$ although the parameter of β is a member of the parameter space B with the dimension q. If $f_\theta \cap g_\beta = \varnothing$, then it is called non-nested. Otherwise, it is nested. There is a need for a parameter to maximize the second part of the Kullback-Liebler information criterion (1) and minimize the measure of information in order to minimize the measures of Kullback-Liebler information criterion, which makes a distinction between the competitive and true models and makes it possible to achieve a desirable competitive model. If the competitive model is well defined to contain the true model, it means that θ_0 belongs to Θ so that $h(x) = f_{\theta_0}(x)$.

Therefore, the maximum likelihood estimator of $\hat{\theta}_n$ is obtained from the derivation of logarithm of the likelihood function for the complete model. As a result, $\hat{\theta}_n$ is converged to θ_0 which is the true parameter of data, and data is generated from it. If the competitive model is mis-specified, it means that it does not contain the true model of data, and the maximum quasi-likelihood considered by White (1982) will be converged to θ_*. In both conditions, the measure of Kullback-Liebler information criterion is minimal. The existing measures of incomplete data are considered because the data is not always complete, as mentioned before. Truncated distributions, therefore, are a conditional distribution resulting from limitations placed up on the probability distribution range. This truncation can be applied from right or left or both sides.

For the truncation purpose, a double truncated distribution is introduced. Consider $X_1, \ldots, X_n$ as an independent random variable with the continuous density function f and the cumulative distribution function F. The logarithm of

quasi-likelihood function for the truncated data from both sides is calculated as follows:

$$Lf_n(\theta)=\sum_{i=1}^{n}\log f_\theta(x_i)I_{(a,b)}(x_i)-\log\left(F_\theta(b)-F_\theta(a)\right), \tag{2}$$

so that $I(x)$ shows indicator function, and the maximum of the quasi-likelihood estimator that holds true in the condition

$$Lf_n\left(\hat{\theta}_n\right)=\sup_{\theta\in\Theta}Lf_n(\theta),$$

so that the obtained $\hat{\theta}_n$ converges to the pseudo-true value

$$\theta_*=\arg\max_{\theta\in\Theta}\left\{E_{\frac{h}{H(b)-H(a)}}\left[\log f_\theta(X)\right]-\log\left(F_\theta(b)-F_\theta(a)\right)\right\},$$

in probability. This parameter shows the closeness of the competitive model to the true model of data. If the competitive model contains the true model of data, the quasi-likelihood estimator is the same as the maximum likelihood estimator.

A Study of the Asymptotic Behavior of Maximum Quasi-Likelihood Estimator Under Double Truncated Observations

Consider the asymptotic behavior of the maximum quasi-likelihood estimator under truncated observations. White's assumption (White, 1982) and definition $n_0=\Sigma_{i=1}^{n}I_{(a,b)}(x_i)$ are considered here. Thus,

$$\frac{1}{n_0}\sum_{i=1}^{n}\log f_\theta(x_i)I_{(a,b)}(x_i)\xrightarrow{P}E_{\frac{h}{H(b)-H(a)}}\left[\log f_\theta(X)\right],$$

where $\xrightarrow{P}$ denotes convergence in probability. According to Slutsky's theorem (Slutsky, 1925), the following convergence relation in this probability holds true:

$$\frac{1}{n_0}\sum_{i=1}^{n}\log f_\theta(x_i)I_{(a,b)}(x_i)-\log\left(F_\theta(b)-F_\theta(a)\right)\xrightarrow{P}$$
$$E_{\frac{h}{H(b)-H(a)}}\left[\log f_\theta(X)\right]-\log\left(F_\theta(b)-F_\theta(a)\right),$$

In order to achieve the asymptotic distribution of $\hat{\theta}_n$ by using Taylor's expansion of $n^{-\frac{1}{2}}\frac{\partial}{\partial\theta}Lf_n(\theta)$ about θ_*:

$$0=n^{-\frac{1}{2}}\frac{\partial}{\partial\theta}Lf_n(\theta)|_{\theta=\hat{\theta}_n}$$
$$=n^{-\frac{1}{2}}\frac{\partial}{\partial\theta}Lf_n(\theta)|_{\theta=\theta_*}+n^{-\frac{1}{2}}(\theta-\theta_*)'\frac{\partial^2}{\partial\theta\partial'\theta}Lf_n(\theta)|_{\theta=\theta_*}+o_p(1),$$

where $o_p(1)$ shows the quantity which converges to zero. Also, it can easily be calculated that

$$n^{-\frac{1}{2}}\frac{\partial}{\partial\theta}Lf_n(\theta)|_{\theta=\theta_*}\xrightarrow{L}N\left(0,I_{ft}(\theta_*)\right),$$

where $\xrightarrow{L}$ denotes the convergence in the distribution, and $I_{ft}(\theta_*)$ is defined as

$$I_{ft}(\theta_*)=E_{\frac{h}{H(b)-H(a)}}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\frac{\partial}{\partial'\theta}\log f_\theta(X)\right]$$
$$-E_{\frac{h}{H(b)-H(a)}}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]E_{\frac{h}{H(b)-H(a)}}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]'.$$

Also

$$-n^{-1}\frac{\partial^2}{\partial\theta\partial'\theta}Lf_n(\theta)|_{\theta=\theta_*}\xrightarrow{P}J_{ft}(\theta_*),$$

so $J_{ft}(\theta_*)$ can be defined as:

$$J_{ft}(\theta_*) = -\left\{ E_{\frac{h}{H(b)-H(a)}}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(X) \right] + \frac{\partial^2}{\partial\theta\partial\theta'}\left[\log\left(F_\theta(b) - F_\theta(a)\right) \right] \right\},$$

where $I_{ft}(\theta_*) < +\infty$ and $J_{ft}(\theta_*) < +\infty$. Hence, the asymptotic distribution of the maximum quasi-likelihood estimator can be obtained as follows:

$$n^{\frac{1}{2}}\left(\hat{\theta}_n - \theta_*\right) \xrightarrow{L} N\left(0, I_{ft}(\theta_*)^{-1} J_{ft}(\theta_*) I_{ft}(\theta_*)^{-1}\right),$$

so if the competitive model is well-specified, therefore $I_{ft} = J_{ft}$, and finally

$$n^{\frac{1}{2}}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} N\left(0, I_{ft}^{-1}(\theta_0)\right).$$

Note that for complete observations, $n^{\frac{1}{2}}\left(\hat{\theta}_n - \theta_0\right)$ has a normal distribution with a zero mean and variance that equals the inverse of the Fisher information.

Generalizing Vuong's Examinations of Double Truncated Observations

Hypothesis testing has two classic approaches: the likelihood ratio test and the Neyman-Pearson test. When the competitive model contains the true model, these tests perform well. However, if the competitive models are not well-specified or are non-nested, then different hypothesis tests are required. Cox (1962) and Vuong (1989) studied such models. Using the results obtained from Vuong's test for complete observation, test the equivalence or closeness of the two competitive models for incomplete double truncated observations. The difference between the functions of quasi-likelihood in the two competitive models of f_θ and g_β is shown as $L_n fg(\theta,\beta)$ so that the following can be obtained:

$$\begin{aligned} L_n fg(\theta,\beta) &= Lf_n(\theta) - Lg_n(\beta) \\ &= \frac{1}{n_0}\sum_{i=1}^{n} \log \frac{f_\theta(x_i) I_{(a,b)}(x_i)}{g_\beta(x_i) I_{(a,b)}(x_i)} - \frac{F_\theta(b) - F_\theta(a)}{G_\beta(b) - G_\beta(a)}. \end{aligned}$$

Using the Central Limit Theorem (CLT), it can be shown that

$$n^{-\frac{1}{2}} L_n fg(\theta,\beta) - n^{\frac{1}{2}} \left\{ E_{\frac{h}{H(b)-H(a)}} \left[\log \frac{f_\theta(X)}{g_\beta(X)} \right] - \log \frac{F_\theta(b)-F_\theta(a)}{G_\beta(b)-G_\beta(a)} \right\} \xrightarrow{L} N\left(0, V^2\right),$$

where V^2 denotes variance of difference between the functions of quasi-likelihood in the two competitive models, and it can be calculated as follows:

$$Var_{\frac{h}{H(b)-H(a)}} \left[\log \frac{f_{\theta_*}(X)}{g_{\beta_*}(X)} \right].$$

For a large sample n, $\hat{v}_n^2$ is an estimator for V^2, defined as:

$$\hat{v}_n^2 = \frac{1}{n_0} \sum_{i=1}^{n} \left\{ \log \frac{f_{\theta_*}(x_i) I_{(a,b)}(x_i)}{g_{\beta_*}(x_i) I_{(a,b)}(x_i)} \right\}^2 - \left\{ \frac{1}{n_0} \sum_{i=1}^{n} \log \frac{f_{\theta_*}(x_i) I_{(a,b)}(x_i)}{g_{\beta_*}(x_i) I_{(a,b)}(x_i)} \right\}^2 .$$

Using Vuong's assumptions (Vuong, 1989) under the null hypothesis, the two competitive models are equivalent or have the same closeness to the true model of data, although one of the models is closer than the other. As for the truncated data, the divergence of the two competitive models can be introduced as follows:

$$H_0 : E_{\frac{h}{H(b)-H(a)}} \left[\log \frac{f_{\theta_*}(X)}{g_{\beta_*}(X)} \right] = 0 \Rightarrow \log \frac{F_\theta(b)-F_\theta(a)}{G_\beta(b)-G_\beta(a)} = 0$$

$$H_0' : E_{\frac{h}{H(b)-H(a)}} \left[\log \frac{f_{\theta_*}(X)}{g_{\beta_*}(X)} \right] > 0 \Rightarrow \log \frac{F_\theta(b)-F_\theta(a)}{G_\beta(b)-G_\beta(a)} > 0$$

$$H_0'' : E_{\frac{h}{H(b)-H(a)}} \left[\log \frac{f_{\theta_*}(X)}{g_{\beta_*}(X)} \right] < 0 \Rightarrow \log \frac{F_\theta(b)-F_\theta(a)}{G_\beta(b)-G_\beta(a)} < 0$$

Therefore, under H_0 hypothesis $n^{-\frac{1}{2}} L_n fg\left(\hat{\theta}_n, \hat{\beta}_n\right) / \hat{v}_n \xrightarrow{L} N(0,1)$, under H_0' hypothesis, $n^{-\frac{1}{2}} L_n fg\left(\hat{\theta}_n, \hat{\beta}_n\right) / \hat{v}_n \to +\infty$, and under H_0'' hypothesis,

$n^{-\frac{1}{2}} L_n fg\left(\hat{\theta}_n, \hat{\beta}_n\right) / \hat{v}_n \to -\infty$. The obtained tests have been run for observation at a significance level of α. If the value of the statistic $n^{-\frac{1}{2}} L_n fg\left(\hat{\theta}_n, \hat{\beta}_n\right) / \hat{v}_n$ is larger than $z_{(1-\alpha)}$, it means that model f is better than model g; if smaller than $-z_{(1-\alpha)}$, model g is better than model f; and if $\left|n^{-\frac{1}{2}} L_n fg\left(\hat{\theta}_n, \hat{\beta}_n\right) / \hat{v}_n\right|$ is smaller than $z_{(1-\alpha)}$ the two competitive models are equivalent. In the defined truncated tests, if $b = +\infty$ and $a = -\infty$, it means that data is complete and is equivalent to Vuong's test. Also, by considering $b = +\infty$ and $a = -\infty$, the obtained maximum quasi-likelihood estimator and related statistical tests for truncated observation from left and right, respectively, are developed.

Simulation Study

Consider the hypothesis testing process for the obtained statistics under truncation from the left side ($b = +\infty$). Independent and identically distributed data is simulated from exponential distribution having parameter 3, which is our true model. The exponential distribution (f) with parameter λ, and Wiebull distribution (g) with parameters α and β, are considered here for the competitive model.

Table 1. Selecting between Exponential and Weibull models using obtained statistical test

	a					
n	**0.01**	**0.03**	**0.05**	**0.10**	**0.20**	**0.30**
10	1.600	3.892	5.971	3.090	6.473	7.065
100	14.310	32.360	40.710	47.230	44.710	26.800
1,000	141.220	209.700	416.900	523.230	542.240	498.450
5,000	442.370	924.290	1,172.940	3,300.520	2,779.760	2,117.980
10,000	497.230	820.530	966.420	3,028.550	4,086.200	4,189.090
100,000	1,981.980	6,076.680	10,341.380	32,165.980	45,621.050	48,387.410

The parameter estimation is conducted as described here. Then the statistical tests are run several times. The results are presented in Table 1, designed for sample size 10, 100, 1,000, 5,000, 10,000 and truncation 0.01, 0.03, 0.05, 0.1, 0.2 and 0.3. Because the obtained test statistics are normal, by enlarging the sample size, they converge to $+\infty$ more accurately. These results confirm the closeness of the exponential model to the true model of data, preferred over Weibull's model.

Conclusion

The characteristics of the maximum likelihood estimators of the truncated distribution were discussed. Some statistical tests were introduced to determine the closeness of the competitive models to the true model for truncated and incomplete data. Under the null hypothesis, these tests turned out to be equivalent. When the null hypothesis was rejected, one of the competitive models was closer to the true model for the data. Results obtained from the current study can be extended to other types of truncation distribution.

References

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In B. N. Petrov and F. Caski (Eds.). *Proceedings of the Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, 1973, pp. 267-281.

Bhattacharyya, G. K. (1985). The asymptotics of maximum likelihood and related estimators based on type II censored data. *Journal of the American Statistical Association, 80*(390), 398-404. doi: 10.1080/01621459.1985.10478130

Bardly, E., & Vahe, P. (1999). Nonparametric methods for the testing and estimation of doubly truncated data. *Journal of the American Statistical Association, 94*(447), 824-834. doi: 10.1080/01621459.1999.10474187

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics and Probability Letters, 44*(4), 333-344. doi: 10.1016/S0167-7152(98)00200-4

Cox, D. R. (1961). Test of separate families of hypothesis. In J. Neyman (Ed.). *Proceedings of the Fourth Berkeley Symposiumon Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 105-123. Berkeley, CA: University of California Press.

Cox, D. R. (1962). Further result on tests of separate families of hypothesis. *Journal of the Royal Statistical Society, Series B (Methodological), 24*(2), 406-424.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79-87. doi: 10.1214/aoms/1177729694

Lominashvili, G., & Patsatsia, M. (2013). On the estimation of a maximum likelihood of truncated exponential distributions. *Bulletin of the Georgian National Academy of Sciences, 7*(1), 21-24.

Slutsky, E. (1925). Uber stochastische asymptoten und gernzwerte. *Metron* (in German), *5*(3), 3-89.

Shimodaira, H. (1994). A new criterion for selecting models from partially observed data. In P. Cheeseman and R. W. Oldford (Eds.). *Selecting Models from Data: Artificial Intelligence and Statistics IV, Lecture Notes in Statistics*, pp. 21-29. New York: Springer. doi: 10.1007/978-1-4612-2660-4_3

Vuong, Q. H. (1989). Likelihood ratio test for model selection and non-nested hypothesis. *Econometrica, 57*(2), 307-333. doi: 10.2307/1912557

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*(1), 1-25. doi: 10.2307/1912526